SOFTWARE REQUIREMENT COLLECTION ENHANCEMENT USING SAMPLING TECHNIQUE AND APPLYING T-DISTRIBUTION

Swarnalatha K S1, G N Srinivasan2, Rakesh R3, Vipin Dwivedi4

RV college of Engineering,
Mysore Road, R V Vidyanikethan Post, Bangalore, Karnataka 560059

ABSTRACT

In software engineering, effective development phase can only takes place if the requirements gathering from client side are clearly defined and should be approachable. Client’s requirements play a crucial role for completion of the successful software projects in all of the development all the software methodologies. If requirements from the client are not properly defined it can lead to failure of the project. Now a day’s some information technology organization are not aware about the proper way to collect the requirement from the clients and fulfill the requirements of the clients. Effective Software development can only takes place if the requirements are properly defined and modeled. Most of the Clients have difficulty in explaining what they need exactly, and the problems increases when software developers failed to convert the client’s requirements into working code. The role of the business analyst to effectively communicate with clients to understand the requirements. Many of project fail due to inappropriate information gathering process.

Index Terms: Requirement Engineering, SDLC, Gathering, Sampling Theory, t-distribution.

1. INTRODUCTION

Software Engineering is the very important and its a integrated part of the any software or hardware industry from past several years. Software industry have become significant efforts based activity from the last few years. Around 99.2% of software development companies which are very small i.e. less than 100 employees in a software company are working towards significant products, for those software firms which needs efficient software requirement engineering practices that are suitable for their defined size and any type of business based the client requirement. Developing and implementing any efficient software is a difficult and extremely software developer intensive activity [1]. Developing and deploying the software is error prone because of many software developer
intensive activities. Day by day software based devices are controlling functions that are very crucial to human survival. The chances of failures of these software based devices have increased rapidly.

**Problem Definition**

Effectively requirement gathering from the clients’ requirements is an important and first step of any software projects and one of the most challenging requirement project management skills. It very important to build the software project on well-defined platform and understandable client requirements to avoid cost burden, unsatisfied client, or even cancellation of the project. Clients typically find difficulty in explaining requirement to the business analyst or project manager[2]. The main problems increase when software developers fail to translate requirements collected from clients into working software program. Main focus of study is on the analysis of client’s requirements gathering or collection practices using sampling theorem.

**Background**

Various studies highlights the requirement collection is vitally important to build the software project on well-formed platform and verifiable client's requirements to avoid cost burden, unsatisfied clients, or even cancellation of the project. Clients typically have difficulty in explaining what they need exactly and the problems rapidly increases when software developers fail to translate requirements from the clients into working software programs. American Standish Group published a survey report called chaos report on software requirement collection. The failure rate of all type of software companies are presented below:

- The success rate was only 24% in large software companies, 37.2% in medium scale software companies and 48% in small scale software companies.
- 69.5 % of large software company projects were very challenging, compared to 52.7% in medium software companies and 60.4% in small software companies.
- 39.5 % projects were cancelled in large software companies, compares to 45.1% in medium software companies and 31.6% in small scale software companies

**2. MATHEMATICAL MODEL**

This sampling theorem is used in estimating population. We are using the same theorem in estimating software requirements [3, 4]. We are estimating the mean for a particular requirement for the entire population i.e., the mean of responses of a huge number of people. Instead of collecting requirements from a very large number of people, we have to just collect the requirements from few people and this collected requirements is a sample. We have to collect few more samples in the similar method. Samples have to be verified before it is taken into consideration (Samples cannot be biased). To employ t-distribution the samples under consideration have to be in normal distribution or binomial distribution [6]. Here we are employing binomial distribution. The questions are of type ‘yes’ or ‘no’ (most preferred since we are using binomial distribution), if not ‘yes’ or ‘no’ has to be assumed internally by the requirement collector[7].

For the development of a food application software for a mobile, a survey was conducted by us out of which one of the question was “Do you like to Cook?” – This was the question which was for analysis of the application’s download [5]. The answers to this question was ‘yes’ or ‘no’.

The mean for the entire population estimated through taking samples will lie within a particular range (range will be calculated) and the mean for the entire population can never go beyond this range (having a confidence level 99% for two sided which means one of the range calculated for 100 entries lies outside the mean’s value.). Let us divide the input collected into
different samples of size 10\([9]\). The following table shows the response of people saying ‘yes’ to the question.

<table>
<thead>
<tr>
<th>Sl.No</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of People saying ‘yes’</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Probability of saying ‘yes’</td>
<td>0.6</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>Probability of saying ‘no’</td>
<td>0.4</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

From binomial distribution
The mean can be calculated by using the formula

\[ \bar{X} = np \]

The Standard Deviation can be calculated using the formula

\[ S = \sqrt{npq} \]

\( n \) = number (here the number of people in the sample)
\( p \) = probability (here the probability of saying ‘yes’)\( q = 1-p \) (probability of saying ‘no’)

For the first sample

(Mean) \( \bar{X} = 0.6 \times 10 = 6 \) and (Std. Deviation) \( S = \sqrt{10 \times 0.6 \times 0.4} = 1.549 \)

Similarly \( \bar{X} \) and \( S \) for all the samples can be calculated and the values are shown below.

<table>
<thead>
<tr>
<th>Sl.No</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{X} )</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>( S )</td>
<td>1.549</td>
<td>1.265</td>
<td>1.449</td>
<td>1.265</td>
<td>0.949</td>
<td>1.449</td>
<td>1.549</td>
</tr>
</tbody>
</table>

After this we can apply the t-distribution knowing the sample mean and standard deviation of the sample the range in which the mean of ‘yes’ over the entire population lies can be estimated. The curve obtained will be of bell shaped with population mean in the center \([10]\). In order to do this first we need to decide on the confidence level let us take confidence level as 99\%. The curve will look similar to as shown below.
From $t$-tables value of $t$ for 99% confidence level for two sided curve is 3.250 (here degrees of freedom for calculating $t$ is $n-1 = 9$)

The formula for range can be calculated using

$$
\bar{X} \pm \frac{txS}{\sqrt{n}}
$$

Where, $\bar{X}$ = Sample Mean, $S$ = Standard deviation sample

$t = 3.250$ (from the table), $n$ = sample size (here it's 10)

Calculating for 1st value,

$$
6 \pm \frac{3.25 \times 1.549}{\sqrt{10}} = (4.408, 7.592)
$$

Similarly for calculating range for all values and tabulating

<table>
<thead>
<tr>
<th>Sl.No</th>
<th>6</th>
<th>1.549</th>
<th>(4.408, 7.592)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>1.265</td>
<td>(6.70, 9.30)</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1.449</td>
<td>(5.511, 8.489)</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>1.265</td>
<td>(6.70, 9.30)</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>0.949</td>
<td>(8.025, 9.975)</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>1.449</td>
<td>(5.511, 8.489)</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1.549</td>
<td>(4.408, 7.592)</td>
</tr>
</tbody>
</table>
These values can be plotted and a curve as in Fig. 1.0 will be obtained, the population mean lies in the center and ranges will be as shown in the diagram, since we have used a confidence level of 99% only one range will lie outside the mean value for 100 samples.

To get a good approximation for the range, we can take average of lower order values and higher order values in the range column in the table 1.3 [11]; after calculating the average, good approximation for the range for population mean can be obtained the range obtained is (5.895, 8.678). So, the population mean lies between this range and these ranges can be averaged to get the population mean having confidence level of 99% and the value is 7.332 this is true for the entire population.

CONCLUSION

From the above method, instead of collecting requirements from lakhs of people (using questionnaire) we can simply collect the requirements from few thousands of people and know the entire population’s effect from small amount of input taken and hence can save lots of time compared to the normal method of questionnaire.

REFERENCES


